L'Arabe dans le Dictionnaire en Ligne d'Aujourd'hui Réflexions autour du Logiciel Linguistique Alexandria

KHALEF El Djouher
(Université d'AlgerII)

Résumé:
Dans le cadre de la compréhension automatique de la langue arabe, nous nous proposons d'explorer les possibilités de manipulation de cette langue dans l'un des logiciels d'aide contextuelle les plus performants au jour d'aujourd'hui: Alexandria.Considérant qu'Alexandria représente à la fois une nouvelle génération de dictionnaires et un vrai commencement pour le web sémantique, nous allons nous intéresser, dans le cadre de ce colloque « TRADETAL », aux problèmes de représentation "cognitive" du lexique arabe et aux relations lexicales en contexte. Ensuite, en testant les limites du logiciel par rapport à la langue arabe, d'abord, en termes d'utilisation (praticité, facilité de recherche, options de l'arabe comparée aux autres langues...) et puis, en termes d'efficacité scientifique (pertinence des définitions, des traductions, des dérivations, des synonymes et antonymes ainsi que des pages web vers lesquelles il renvoie), nous rendrons compte des lacunes de l'arabe dans le dictionnaire en ligne actuel et nous tenterons d'apporter des suggestions dans la perspective d'un prototype "amélioré" et "plus adapté" pour l'utilisateur de la langue arabe. Alors, en parlant du dictionnaire pour "ordinateur", les applications de la langue arabe, sont-elles prometteuses ? C'est du moins, ce que nous nous attellerons à démontrer au cours de notre intervention.

1. Introduction et définitions générales

Le Traitement Automatique du Langage Naturel (TALN) ou Traitement Automatique des Langues (TAL) est une discipline à la frontière de la linguistique et de l'informatique, qui concerne l'application de l'informatique au

langage humain. Longtemps, des spécialistes ont travaillé le jeu des relations inter-lexicales. Ces tâches ont abouti aux outils opérationnels suivants:
- LE DICTIONNAIRE INTEGRAL (LDI) qui est un ensemble de ressources linguistiques.
- LE SEMIOGRAPHE qui est un jeu d'APIs (Application Programming Interface) en Java : correction phonétique, étiquetage morphologique, expansion de requêtes, dictionnaire à l'envers, extraction des thèmes, routage et résumé de textes à base d'isotopies.

Aujourd'hui, du point de vue du dictionnaire pour "ordinateur" et plus récemment, du dictionnaire en ligne, le problème serait de trouver la structure informationnelle qu'une occurrence de mot doit créer.

- LE WORDNET est un dictionnaire informatisé dont l'unité de base est le concept non le mot. La conception et la méthodologie du wordnet ont été développées par des linguistes du laboratoire des sciences cognitives de l'université de Princeton pour l'anglais et d'EuroWordnet pour les langues européennes. Il utilise deux moyens pour définir le sens d'un mot:
- Les synsets (synonym set) : un groupe de mots interchangeables, dénotant un sens ou un usage particulier, c'est-à-dire des ensembles de synonymes et de pointeurs décrivant des relations vers d'autres synsets. Chaque mot peut appartenir à un ou plusieurs synsets, et à une ou plusieurs catégories du discours. Ces catégories sont au nombre de quatre : nom, verbe, adjectif et adverbe.

À l'instar d'un dictionnaire traditionnel, WordNet offre ainsi, pour chaque mot, une liste de synsets correspondant à toutes ses acceptions répertoriées. Mais les synsets ont également d'autres usages : ils peuvent représenter des concepts plus abstraits, de plus haut niveau que les mots et leurs sens, qu'on peut organiser sous forme d'ontologies. Une ontologie est un système de catégories permettant de classifier les éléments d'un univers.

- Les relations lexicales : un système de catégorisation correspondant aux relations sémantiques. Il permet de regrouper de manière cohérente toutes les composantes d'un univers linguistique telles que les mots, les sens ou bien les concepts. Les relations suivantes sont utilisées dans wordnet : synonymie, antonymie, hyponymie, méronymie, morphologie.

Connaitre la place d'un mot dans ce réseau de relations, c'est connaitre son sens. Wordnet est donc un réseau lexical dont les synsets sont les nœuds et les relations entre synsets sont les arcs.

Le Wordnet Arabe autour duquel s'articule notre thématique générale est une base de données lexicale librement disponible pour l'arabe standard. Sa structure est celle d'un thésaurus (index alphabétique de mots reliés entre eux par des relations sémantiques), tel que nous venons brièvement de l'expliquer.

Il faut noter toutefois que Wordnet Arabe est une des rares ressources pour la langue générale arabe disponible en ligne. Il compte actuellement 111269 synsets et 23 481 mots.

L'arabe, par sa forme d'écriture (de gauche à droite, signes diacritiques optionnels, pas de majuscule, jusqu'à quatre variantes pour la même lettre...), sa phonologie (vingt-huit consonnes, trois voyelles courtes et trois voyelles longues...), sa morphologie (dérivation à partir de racines, existence d'infixes en plus des préfixes et suffixes, agglutination...) et sa syntaxe (assez libre, absence possible des désinences...) est encore considéré parmi les langues difficiles à appréhender sur le plan du traitement automatique. Les problèmes de recherche que posent les quelques caractéristiques que nous venons d'évoquer, nécessitent une formalisation assez poussée et des outils robustes d'analyse.

Nous allons donc explorer ces facettes de l'arabe dans la version 2 familiale d'Alexandria qui utilise le wordnet.

2. Alexandria : présentation et mode d'utilisation

Alexandria appartient à plusieurs familles de logiciel. Une première famille est celle des logiciels d'aide contextuelle ;en effet, Alexandria fournit un service d'aide à la compréhension (définitions, traductions...) quand un lecteur le sollicite. D'autre part, Alexandria appartient à la famille des agents "intelligents" : ce service est disponible en chaque lieu où l'agent est installé. Aujourd'hui, les lieux où l'on trouve les réplications les plus courantes de cet agent sont les *pages html* du web.

C'est donc un dictionnaire en ligne multilingue, basé sur un réseau sémantique considérable, traitant dans ses dernières versions 22 langues interrogeables entre elles, issu de la recherche française et développé par la société Mémodata créée en 1989 par *Patrick de Torcy* et *Dominique Dutoit* qui a eu

l'idée d'Alexandria. Il est téléchargeable en version d'évaluation gratuite et s'enrichit en permanence de vocabulaires spécialisés.

Alexandria est un hypermédia multi source contextuel personnalisable qui peut ouvrir des contenus à l'intérieur de lui-même ou non. "Couteau suisse" des dictionnaires, il offre à ses utilisateurs d'un clic de souris l'ouverture contextuelle des informations linguistiques propres à leur langue ou aux langues de leur document. Permettant l'accès aux plus grands réseaux sémantiques actuels, il s'introduit dans le web sémantique d'une façon directe en facilitant le passage de l'idée à la découverte d'un mot qui peut précéder une interrogation sur les moteurs de recherche ; et d'une façon indirecte, en facilitant le passage des termes aux termes conventionnels et logiques que le futur web sémantique met progressivement en place. Selon leurs besoins, les utilisateurs d'Alexandria bénéficient de ces deux directions pour rendre plus accessible ou plus intelligible les termes propres à leurs documents.

Plus encore, l'application web valorise Alexandria qui conduit à un mode d'applications très diversifiées. Le concepteur du dictionnaire cite l'exemple suivant : « *Je suis réparateur d'avions et mon collègue m'envoie un mail me disant qu'il faut appliquer la procédure MT-25-4... Je n'irai jamais dans mon moteur de recherche pour vérifier cette procédure MT-25-4 si je crois m'en rappeler vaguement. Tandis qu'avec ce que propose Alexandria (il suffit de cliquer sur le mot pour avoir tout de suite le résultat), les langues de spécialités (MT-25-4) et les langues générales finissent par converger. On est ainsi toujours en situation d'apprenant, mais en plus, encouragé par le système à être cet apprenant-là.* »

(Interview de *Dominique Dutoit* in Automates Intelligents, 2006). Le concepteur visait le développement d'un dictionnaire du français qui augmenterait de façon sensible les liens analogiques entre les mots, de telle façon que la recherche de termes précis, pertinents, soit facilitée. « *L'idée générale, toujours en grande partie vraie aujourd'hui étant d'apporter une sorte de réponse à la célèbre phase de Boileau "Ce qui se conçoit bien s'énonce clairement et les mots pour le dire viennent aisément" : ce qui est dit avec facilité montre une belle maîtrise des mots, et le discours clair favorise une bonne conception. Il s'agit de faciliter l'expression par une meilleure transmission des vocabulaires, précis ou riches*. » (Op.Cit).

Alexandria possède les caractéristiques principales suivantes:

- C'est une encyclopédie électronique avec tous les services que cela comporte : synonyme, définition, noms propres, analogie ("des idées vers les mots").

- C'est aussi un dictionnaire électronique sémantique. Une de ses principales forces est de pouvoir ajouter une langue (exemple, le Chinois) en "traduisant" uniquement le Chinois en Français. Il se charge ensuite de faire la traduction automatique de cette nouvelle langue dans toutes les autres existant déjà dans Alexandria.

- Alexandria permet aussi, grâce à une recherche souple et intuitive, l'accès aux grandes quantités d'informations (textes, images, données techniques, ...) de l'entreprise (Intranet) ou sur le Net.

- Le moteur utilisé pour cette encyclopédie est aussi utilisable pour gérer et accéder à de grandes quantités d'informations,

non nécessairement linguistiques et non nécessairement structurées (textes, images, données techniques, nomenclatures, ...).

Contenu de *Alexandria* édition familiale :

Alexandria édition familiale est considéré comme un puissant outil de productivité intellectuelle à usage personnel et familial. D'une façon générale, en plus de fournir d'excellents dictionnaires, il s'agit du premier navigateur capable d'enchaîner des recherches par mots clés sur tous les moteurs de recherche.

Par exemple, en recherchant « Beethoven », il est possible, sans refaire cette saisie, d'effectuer la recherche de ce compositeur dans *Ebay*, puis dans Amazon, cela peut-être afin de commander une de ses œuvres, ensuite, toujours sans ressaisir, dans *Deezer*, pour l'écouter, sur *Youtube*, pour voir un orchestre le jouer, sur *wikipedia* ou *sensagent* pour retrouver sa biographie ou/et trouver l'ensemble de ses œuvres, etc.

Il peut arriver aussi, en surfant sur le web, de vouloir conserver certains extraits de document pour constituer un dossier. Cela est par exemple typiquement utile durant les rédactions de mémoire ou de thèse pour organiser des citations selon différents mots clés. La fonctionnalité de prise de note rapide aide à conserver ces fragments et permet de les organiser par mot clé et date tout en fournissant la possibilité de leur ajouter un titre.

S'agissant de la traduction, Alexandria permet de définir jusqu'à trois langues de travail. Pour chacune d'elles, l'utilisateur peut accéder aux dictionnaires monolingues et

bilingues concernant ces langues mais aussi directement au menu contextuel de traduction de phrases.

3. Fonctionnement d'Alexandria par rapport à la langue arabe :

La page d'accueil du dictionnaire offre *un cadre de saisie de l'entrée* tout à fait similaire à celle du moteur de recherche Google. L'utilisateur choisira en bas *la langue du dictionnaire* qu'il souhaite consulter puisque cette dernière est *le français* par défaut : nous pointerons pour notre part l'arabe. En troisième lieu, il faut choisir entre les cinq *types de dictionnaires* qu'offre Alexandria:

- *Dictionnaire*, pour effectuer sa recherche dans un dictionnaire monolingue,

- *Traducteur*, et dans ce cas, il y a lieu de pointer la langue-cible parmi la liste des langues qui s'affiche pour voir les traductions, les dérivés, les locutions et le réseau sémantique bilingue de l'entrée,

- *Anagramme,* pour consulter les mots qui peuvent être formés à partir de l'entrée par renversement de lettres,

- *Conjugueur,* pour la conjugaison des verbes, flexions des noms (pluriel) et des adjectifs (féminin, pluriel).

- *Joker ? et* *, ce dictionnaire effectue une recherche avec jokers. '?', qui remplace une lettre, et '*', remplaçant plusieurs lettres. Mais il n'est disponible qu'en anglais, français, italien, espagnol, portugais, allemand et néerlandais.

Pour notre test, nous insérons dans la case de saisie des mots l'entrée : □ كتب □ , nous sélectionnons la langue arabe et nous pointons sur *Dictionnaire* pour voir les définitions, les synonymes, les dérivés, les locutions et le réseau sémantique que propose Alexandria. Nous lançons la recherche et le dictionnaire donne d'abord les antonymes رتجل et منطوق. Pour les synonymes de □ كتب □*(adj.) :* معلّب puis, □ □ كتب*(v.),* nous avons la liste suivante :

ألّف، أمر، حدّد، حرّر، قاعدة، وصف، يَصِفُ الطبيبُ الدّواءَ، يَكْتُب، يَكْتُبُ رِسالَةً، يُؤَلِّف، يَكْتُبُ كِتاباً أو قَصيدَةً.

Concernant les locutions formées avec notre entrée, Alexandria propose cette liste :

بائِعُ كُتُب • خَزانَةُ كُتُب • دودَةُ كُتُب ، مولَعٌ بالقراءة • رَفُّ كُتُب • كتب السّنويّة • كتب بأحرف الرومانية • كتب ذات الأغلفة الورقية • كُتُبٌ مُعْتَرَفٌ بِحَقيقَتِها • مجلِّد كُتب • مُجَلَّد يَضُم مَجموع كُتُب المُؤلِّف • مُخَرْبِش، مؤلِّف كُتُب ومقالات تافِهَه

أهم كتب السيرة النبوية • أهم كتب النحو • الكافي (كتب • (بحث كتب جوجل • قائمة كتب السحر العربية • قائمة كتب روحانية • قائمة كتب شيعية •قائمة كتب يوسف القرضاوي • قائمه كتب شيعيه • قائمه كتب يوسف القرضاوي • كتب ابن عربي • كتب الأحاديث • كتب التفسير • كتب الحديث •كتب الحديث التسعة • كتب الحديث الستة • كتب السنة التسعة • كتب السيرة • كتب الصحاح • كتب الكترونية • كتب تفسير القرآن • كتب روحانية • كتب ستّة • كتب ستّه • كتب سماوية • كتب سماوية إسلامية •

كتب غير قانونية مشار إليها في الكتاب المقدس • كتب غير قانونيه مشار إليها في الكتاب المقدس • كتب فرعونية • كتب مقدسة • كتب مقدسه • كتب هندوسية • كتب يوسف القرضاوي • كتبٍ • ويكي كتب

Le dictionnaire analogique propose également la définition de notre entrée **كتب** par synonymie et antonymie avec les mêmes résultats suscités. Ensuite, par rapport à la langue française et anglaise, il donne l'hyperonyme et le dérivé de l'entrée comme suit:

- Writing(en) [Domaine]
- Ecrire (fr) [Domaine]
  - أشعر, أوصل, نقل, يَتَّصِل [Hyper.]
  - حرّر, كتب [Hyper.]
  - كاتب - كتابة [Dérivé]

En fin de page de recherche, Alexandria nous renvoie vers Wikipedia, l'encyclopédie libre sur internet dont l'icône se trouve sur la marge gauche de la page des résultats aux cotés de celle du dictionnaire Français-Anglais par défaut, de Google, de eBay (France) et d'une icône supplémentaire pour effectuer d'autres recherches par suggestion.

A présent, nous formulons notre requête pour la même entrée **كتب** dans le *Traducteur* d'Alexandria.

- Pour le français comme langue-cible, voici le résultat :

كتب*(v.)* à, correspondre, écrire, mettre par écrit, prescrire, rédiger, tracer sur une surface.

- Pour l'Anglais, nous avons :

كتب

scripted

كتب *(adj.)*
canned, transcribed, written
كتب *(n.)*
books
كتب *(v.)*
compose, dictate, get down, indite, order, pen, prescribe, put, put down, set down, write, write down

Discussion générale

Le mode de fonctionnement du wordnet arabe dans notre dictionnaire en ligne « Alexandria » consiste à récupérer une liste des termes reliés à l'entrée (**كتب** pour notre exemple) par des relations de synonymie, d'antonymie, de dérivation, et d'hyperonymie. L'entrée est également conceptualisée dans des locutions.

Cependant, les résultats obtenus posent le problème des variations lexicales. Même si des relations sémantiques existent entre l'entrée et les mots proposés par le dictionnaire, il reste le souci du choix du sens (synset) à prendre dans le cas de la polysémie de l'entrée. Ainsi, les locutions proposées pour notre exemple vont beaucoup plus dans le sens de « livres » كُتُبque dans celui du verbe « écrire » كَتَبَ. Ce problème est d'autant plus accentué par la nature grammaticale de l'entrée qui conditionne la pertinence de la requête. Le wordnet arabe d'Alexandria n'utilisant pas un nombre suffisamment grand de relations sémantiques, le sens des mots n'est pas réellement défini car chaque synset est pourvu d'une glose décrivant son sens.

Nous noterons néanmoins que le dictionnaire nous fournit un tant soit peu de savoir syntagmatique, c'est-à-dire, l'information sur les contextes dans lesquels l'entrée et ses acceptions apparaissent.
En fin de résultats, l'utilisateur est renvoyé vers les pages web de Wikipedia ou Google pour l'enrichissement de sa requête.

Du coté du *Traducteur* d'Alexandria, nous pouvons de prime abord remarquer que les résultats sont plus riches en Anglais langue-cible qu'en Français. En effet, alors que la traduction arabe/français ne nous indique que les équivalents de notre entrée □كتب□ en tant que verbe à l'infinitif, la traduction arabe/anglais propose des équivalents de □كتب□ en tant que verbe au passé, adjectif, nom et verbe à l'infinitif. Ceci nous confirme bien que le wordnet est plus performant en Anglais que dans les autres langues où il reste encore du travail à faire.

Le point fort d'Alexandria est certainement de traiter la requête de manière à optimiser le temps de recherche en offrant plusieurs types de dictionnaires et en intégrant les liens internet pour une recherche approfondie. Mais comme pour toute recherche automatique, l'obtention de résultats bons et précis dépend de la formulation de la requête.

L'exactitude de la formulation de la requête en langue arabe est une tâche très délicate eu égard à ses spécificités d'écriture de droite à gauche, son alphabet de 36 lettres et son script, ses signes diacritiques optionnels mais qui participent grandement dans la détermination sémantique, sa morphologie, sa typographie etc. Ainsi, plusieurs options du logiciel linguistique « Alexandria » ne sont pas actives pour la langue

arabe dont les dictionnaires des *anagrammes*, *le conjugueur* et *le joker ? et* *. Tout cela demande par conséquent, beaucoup d'investigations.

4. **Conclusion**

Le logiciel linguistique Alexandria sur lequel s'est basée notre étude de la langue arabe dans le dictionnaire en ligne d'aujourd'hui nous a fourni des résultats contextuels avec les signes diacritiques propres à l'Arabe. Deux acquis indéniables de la bonne application du wordnet arabe dans ce dictionnaire. Nous concluons également avec certitude que la qualité de la réponse que nous obtenons dépend largement de la qualité de la requête construite. Mais il reste que la formulation claire de la requête est bien plus difficile et complexe que la réponse en soi. Une amélioration possible du logiciel consisterait à palier le problème des variations morphologiques et lexicales en général par l'exploitation des formes de base des mots de la requête (ou des entrées du dictionnaire).

Beaucoup de tests et d'améliorations restent à faire quant au traitement automatique de la langue arabe et comme perspective, des chercheurs sont entrain de construire un corpus de textes arabes avec lequel ils pensent faire une évaluation objective de l'apport réel de cette approche.

Références

ABDERRAHIM, M. El A., & al., *Un modèle objet pour le traitement automatique de l'arabe voyellé ou non*, JeTIC'2007, Bechar 21/22 avril 2007.

BESSOU, S. SAADI, A. et TOUAHRIA, M., *Vers une recherche d'information plus intelligente : application à la*

langue arabe, 1ère Conférence Internationale : Systèmes d'Information et Intelligence Economique SIIE 2008 Hammamet, Tunisie, 14-16 Février 2008.

CHAUMARTIN, F-R., *WordNet et son écosystème : un ensemble de ressources linguistiques de large couverture*, BDL-CA 23 avril, Montréal, Canada.

HABASH, N-Y.(2010) *Introduction to Arabic Natural Language Processing*, Morgan & Claypool Publishers, University of Toronto.

MUSA, Alkhalifa.(2006) *Arabic WordNet and Arabic NLP*, JETALA 5-7 June, Rabat, Maroc.